Big Data Applications in TURKSTAT

Digital Transformation and Projects Department

Big Data Group

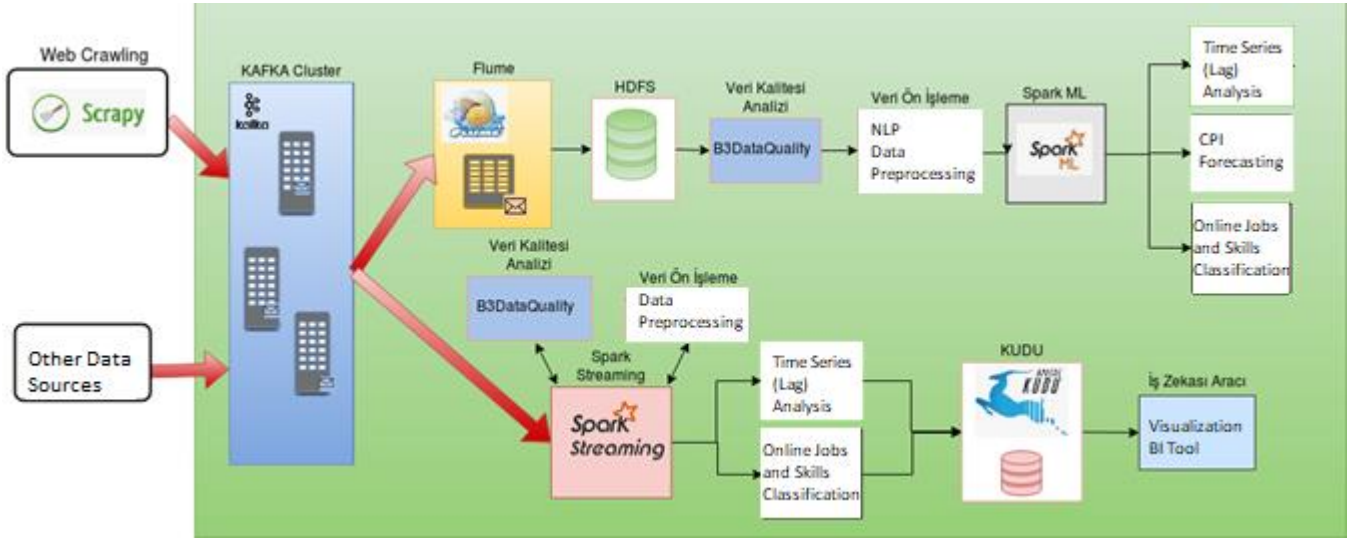
Outline

- **TurkStat Big Data Group**
- **Project 1: Enhancing Performance of Rule-based COICOP Assignation by Using Big Data Tools**
- **Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach**
  - Background
  - Data collection & deduplication phase
  - Assigning COICOP with ML Techniques
  - Assigning COICOP with DL Techniques
- **Project 3 : Assigning ISCO codes to Job Advertisements**
  - Background
  - Data collection
  - Assigning ISCO codes with ML Techniques
  - Assigning ISCO codes with DL Tecnniques

TurkStat Big Data Group

Several project in development

- Working on big data infastructure
- Using machine learning/deep learning algorithms
- Based on different data sources (web scraping, scanner data)
- Both batch processing and stream processing



TurkStat Big Data Group

Projects consists of these stages

- Collection of daily product and online job data from up to 90 web sites
- Preprocessing, deduplication and classification of daily products by using Natural Language Processing (NLP)
- Developing machine learning and deep learning models to perform time series analysis to identify the effects of prices changes between product groups in terms of time interval and ratio.
- Developing machine learning and deep learning models to classify jobs and skills based on online job vacancy data.
- Applying developed models on daily collected data to perform lag analysis on product groups in CPI basket.
- Applying developed models on daily collected online job vacancy data to visualize jobs and skills demands based on sectors, locations and times.

Project 1: Enhancing Performance of Rule-based COICOP Assignation by Using Big Data Tools

- Previously implemented via Python programming language in a traditional way
- 150K code assignation daily with this way
- Re-implemented with Apache Spark runs on 16-nodes Hadoop cluster

Project 1: Enhancing Performance of Rule-based COICOP Assignation by Using Big Data Tools

- 4M code assignation daily with this approach
- Thanks to this approach code assignation process which takes 200 days previously can be completed in 7 days (for 30M rows dataset).
- This time can be further reduced by adding new nodes to the existing Hadoop cluster.

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Background

- Rule-based code assignation processes are not maintainable

- When new products are received, new rules may be required.

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Dataset

- There are 3 main sources.
  - Web scraping data
  - Scanner data
  - Survey data (from Turkstat regional offices)

- Scanner data and survey data are labelled

- Main Columns: Product definition, COICOP code (if assigned)

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Dataset

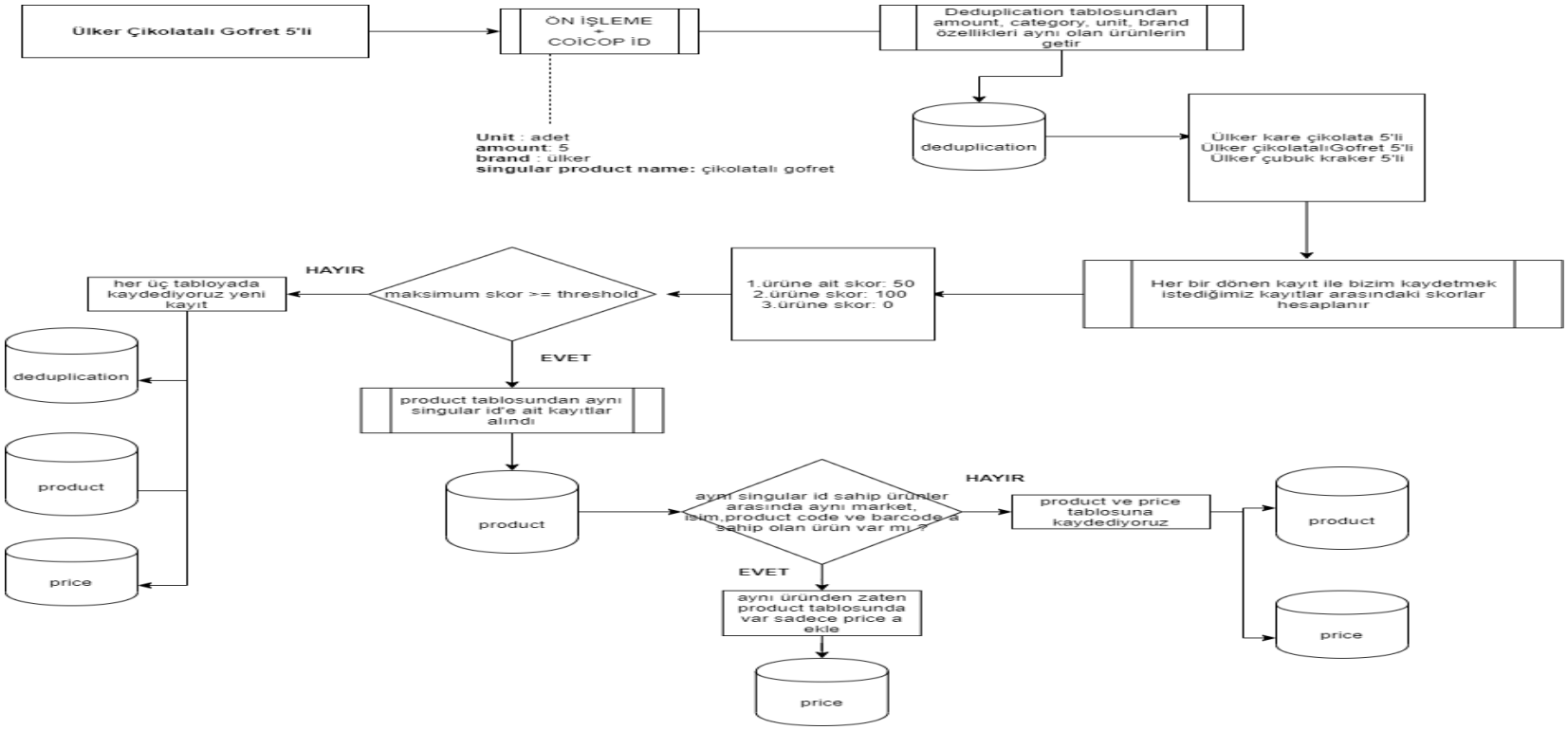
The problem is how to save web scraped data to the cluster.

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Related Spark Code

```
ssc = StreamingContext(sc, 2)
price_kvs = KafkaUtils.createDirectStream(ssc,
                                          config['price_topics'],
                                          config['brokerList'],
                                          valueDecoder= lambda x: x)
price_lines = price_kvs.map(lambda x: (x[1]))
price_lines.foreachRDD(doNecessary4Price)
ssc.start()
ssc.awaitTermination()
```

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach



Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

- Spark codes reads data from Kafka topics, do some preprocessing jobs on streaming data and save to Hive tables.

web_site	product_name	product_price	product_code	category	normalized_extra_info	singular_product_name	unique_product_name	amount	unit	brand	coicop_id
cagri hipermarket	Koroplast Doğada Çözülebilen Çöp Torbası 10lu	9,95	10098	market	koroplast.adet	dogada cozulebilen cop torbasi	koroplast dogada cozulebilen cop torbasi 10	10	adet	koroplast	5612030401
cagri hipermarket	Knorr Kremalı Mantarlı Makarna Sosu	4,95	1773	market	knorr.adet	kremali mantarli makarna sosu	knorr kremali mantarli makarna sosu	1	adet	knorr	111501
cagri hipermarket	Torku Süt Kakaolu 6x180 ml	8,95	7644	market	torku sut.adet	süt kakaolu	torku sut kakaolu 1080.0 mililitre	1080	mililitre	torku sut	114101
cagri hipermarket	İçimino Çikolatalı Süt 6x200 ml	7,95	3924	market	icim.adet	cikolatali sut	icimino cikolatali sut 1200.0 mililitre	1200	mililitre	icim	114101
cagri hipermarket	Tamek Reçel Böğürtlen 380 gr	11,45	10368	market	tamek.adet	recel bogurtlen	tamek recel bogurtlen 380.0 gram	380	gram	tamek	118201
cagri hipermarket	Danone MilkShake Çilek-Vanilya 220 ml	4,25	4566	market	danone.adet	milkshake cilek vanilya	danone milkshake cilek vanilya 220.0 mililitre	220	mililitre	danone	116110

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Machine Learning Approach

- scikit-learn library has been used.
- Some preprocessing steps have been experimented. (TF-IDF, lemmatazion etc.)
- Scikit-learn codes were executed on local computers. (by fetching data from cluster to a local computer)

Model	Accuracy	Precision	Recall	F-score
Logistic Regression	0.96	0.94	0.93	0.93
Support Vector Machine	0.94	0.90	0.90	0.90
Naive Bayes	0.92	0.83	0.79	0.80

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Machine Learning Approach (What's next)

- sparknlp and spark.ml libraries are being planned to use.
- Ensemble learning models such as XGBoost may be used in next.

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Deep Learning Approach

- 3 HuggingFace models have been tried. (dbmdz/bert-base-turkish-cased, dbmdz/bert-base-turkish-128k-uncased and bert-base-multilingual-cased)

- Codes are implemented via Pytorch and executed in Google Colab environment. (for utilizing GPU)

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Deep Learning Approach

MODEL	Precision	Recall	F1-Score
dbmdz/bert-base-turkish-cased	0,91	0,92	0,91
bert-base-multilingual-cased	0,91	0,91	0,91
dbmdz/bert-base-turkish-128k-uncased	0,93	0,92	0,93

Project 2: COICOP Assignation to Web Scraped Products with ML-based Approach

Deep Learning Approach (What's next?)

- Using product images that are collected by pollsters, we are planning to fine tune a pre-trained transfer learning models.

- After we constitute our models, we are planning to apply some ensemble learning techniques by utilizing both our text-based and image-based models.

Project 3: Assigning ISCO Codes to Job Advertisements

Background

- Using survey and administrative data may be both time-consuming and require human force.
- These data may not reflect the market demand at that time.
- Our purpose is to build the ml-based ISCO assignment model with labeled training dataset and use this model on daily web scraped data.

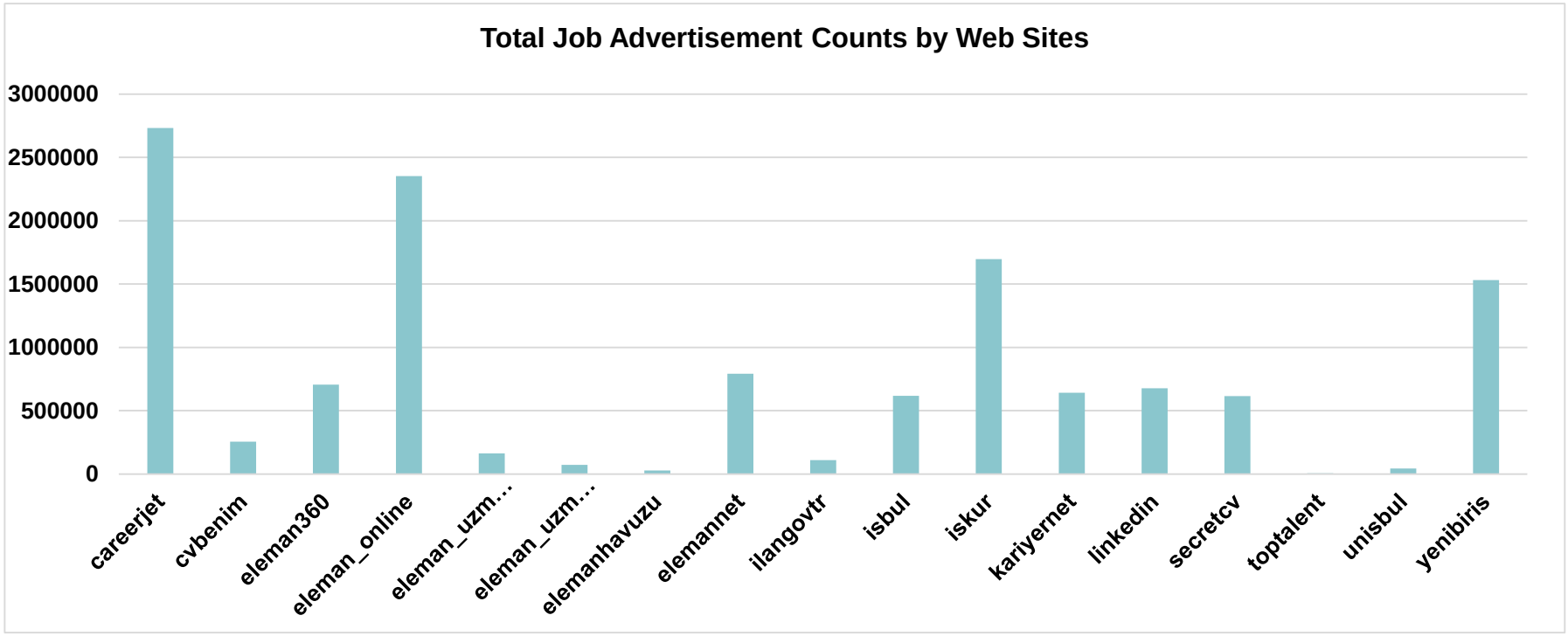
Project 3: Assigning ISCO Codes to Job Advertisements

Data

- 16 different web sites
- Approximately 40K job advertisement daily
- Some columns : title, occupation, city, company name, date_poster, gender etc..

Project 3: Assigning ISCO Codes to Job Advertisements

Data



Project 3: Assigning ISCO Codes to Job Advertisements

Machine Learning Approach

- The machine learning model was trained by using Apache Spark ml library on big data cluster.
- Some Natural Language Processing preprocessing methods (such as N-gram, TF-IDF etc.) have been tried.
- Support Vector Machines and Logistic Regression models have been used at this stage.

Project 3: Assigning ISCO Codes to Job Advertisements

Machine Learning Approach

	Class count	Job adv. Count for each ISCO code	F1-score	# of rows
SVM	32	>100	%88.4	6501
Logistic Regression	32	>100	%83	6501
SVM	77	>50	%82.91	9616
Logistic Regression	77	>50	%75	9616
SVM	127	>25	%76.70	11408
Logistic Regression	127	>25	%65	11408
SVM	210	>10	%62.31	12662
Logistic Regression	210	>10	%57	12662
SVM	354	>=1	%56.94	13224
Logistic Regression	354	>=1	%56	13224

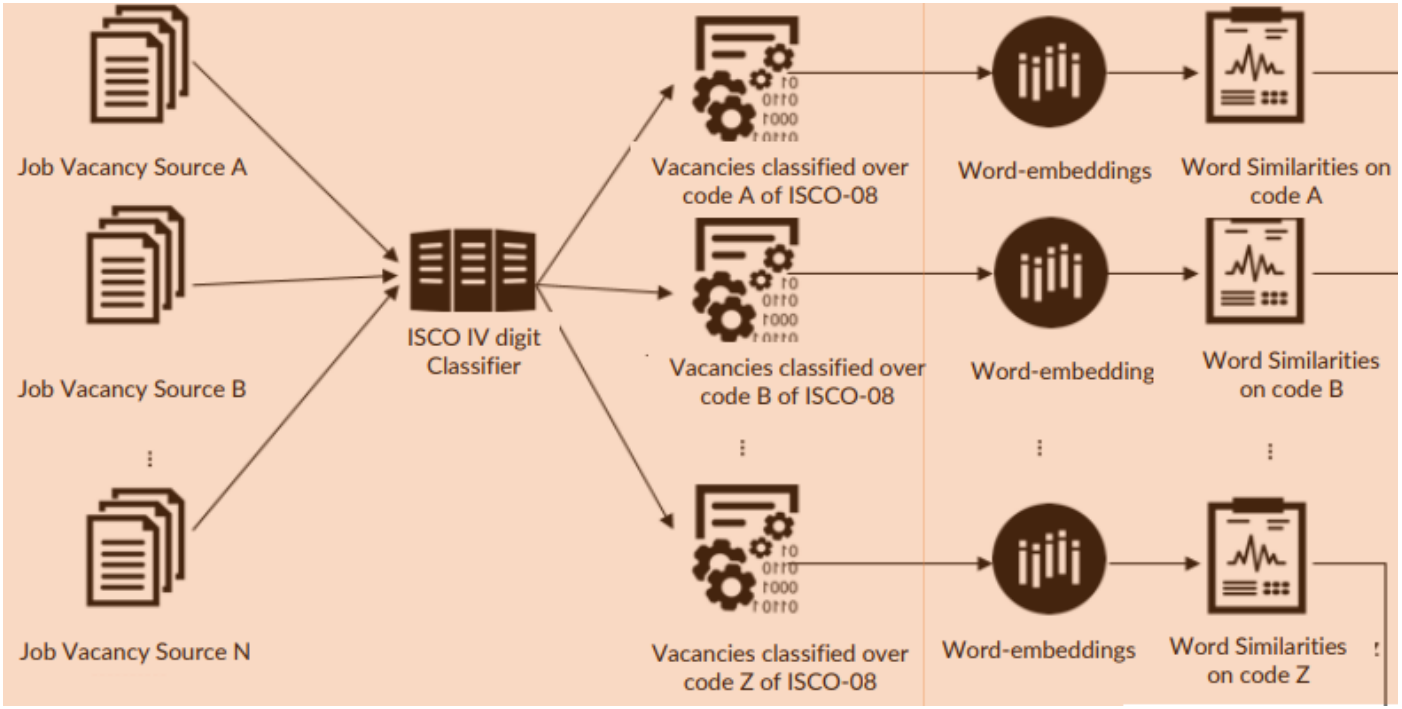
Project 3: Assigning ISCO Codes to Job Advertisements

Deep Learning Approach

- Some pre-trained models on HuggingFace have been fine-tuned with our dataset on Google Colab environment.
- We have tried different approaches like Word2Vec, BERT etc.
- The best trained model in terms of accuracy have been selected and downloaded to our local system and it is ready to make predictions.

Project 3: Assigning ISCO Codes to Job Advertisements

Deep Learning Approach



Project 3: Assigning ISCO Codes to Job Advertisements

Deep Learning Approach

Class count	Job adv. Count for each ISCO code	F1-score
22	>=1000	%92
76	>=500	%83
154	>=250	%76
222	>=150	%71
276	>=100	%69
350	>=50	%64

Thank you for your attendance